International Journal of Computer Applications Technology and Research
Volume 14–Issue 11, 94 – 117, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1411.1008

The Role of Predictive Analytics in Shaping Public Policy for Preventive Healthcare in the United States

Rhoda Kalu Tasie[1], Daniel O. Adelakun[1], Sylvia Etim[1], Adetola Odebode[1]

[1]University of Arkansas, Fayetteville, Arkansas, United States

Abstract

The United States carries the highest chronic disease burden among high-income countries, yet preventive services reach the populations that need them least often. This study examined these relationships and assessed whether survey-based predictors could support individual-level classification and inform preventive-healthcare policy. Data were drawn from the 2024 National Health Interview Survey Sample Adult File (N = 32,629; complete-case analytic samples of 29,110–29,194). A four-stage design used multinomial and binary logistic regression to examine two healthcare utilization outcomes and five self-reported chronic conditions: hypertension, diabetes, cardiovascular disease, high cholesterol, and cancer. Extended models incorporated visit recency and wellness-check status, while Random Forest classifiers provided a complementary assessment of predictive performance. Models were evaluated on held-out test data against majority-class baselines. Continuity of care and visit recency consistently associated with utilization and reported diagnoses, while their associations with insurance coverage weakened after accounting for visit recency, with racial and educational associations varying by condition. Classification accuracy exceeded majority-class baselines for hypertension (+11.3 percentage points) and high cholesterol (+3.4 percentage points), but not for diabetes, cardiovascular disease, or cancer. These findings demonstrate the value of survey-based analysis for identifying population-level patterns in healthcare access and diagnosed disease, while highlighting limits to individual-level prediction using the selected variables. Policy implications include strengthening continuity of care and investigating whether lower reported diagnosis rates among populations facing access barriers reflect unmet screening needs. Cross-sectional, self-reported data limit causal inference and cannot establish undiagnosed disease prevalence.

Keywords: predictive analytics; preventive healthcare; chronic disease; healthcare access; health policy; National Health Interview Survey; logistic regression

1. Introduction

The healthcare system in the United States is dealing with an increasing number of chronic and preventable diseases, even though it spends more per person on advanced technologies and innovations than any other high-income country (Tikkanen & Abrams, 2020; Blumenthal et al., 2024). Recent national reports show that nearly 60% of U.S. adults live with at least one chronic condition, and 40% manage two or more at the same time (Dahima, 2025; Benavidez et. al, 2024). Many of these conditions, such as heart disease, diabetes, high blood pressure, cancer, and chronic obstructive pulmonary disease, are no longer just medical issues but are connected epidemics that affect individuals, families, healthcare institutions, and society overall (Hacker, 2024). According to the Centers for Disease Control and Prevention (CDC, 2026), chronic diseases and mental health conditions make up about 90% of the nation's total healthcare costs, which total nearly $4.1 trillion every year. When indirect costs, like lost productivity in the workforce, are included, the Milken Institute estimates that the total economic burden of chronic disease in the U.S. is roughly $3.7 trillion annually (Waters & Graf, 2018; Raghupathi & Raghupathi, 2018). This is expected to rise to around

$47 trillion by 2030, with individual lifestyle choices, behaviors, and community influences playing a key role in the development and management of chronic diseases (Hacker, 2024; Bloom et al., 2011).

This paper uses logistic regression on the 2024 National Health Interview Survey (NHIS) adult dataset to examine how predictive analytics can help identify the structural factors that lead to chronic disease diagnoses for five conditions: hypertension, diabetes, cardiovascular disease, high cholesterol, and cancer. The main argument is that predictive analytics does more than just identify who is sick; it also maps the conditions under which chronic diseases are not detected or treated, measures the impact of structural access barriers, and identifies which populations are most affected by the current healthcare system. This is important, as it shifts public policy from reactive treatment models toward proactive and equitable preventive care. An additional prospective application of population-health analytics is healthcare resource planning. Patterns of diagnosed chronic disease and care utilization may inform estimates of future service needs; when paired with independently established clinical device-use rates and supply-chain information, such estimates could support medical-device procurement and resilience planning.

Beyond identifying which structural and demographic factors are statistically associated with chronic disease diagnosis, this study also directly tests whether those same factors are sufficient to predict diagnosis at the individual level, evaluating each model's classification accuracy against a naive baseline. This distinction between statistical association and predictive accuracy is treated as a substantive finding in its own right, clarifying where structural survey data can most effectively inform policy and where their usefulness reaches a limit..

2. Literature Review

2.1 Chronic Disease Burden and the Underuse of Preventive Care

Despite many years of public health efforts, the delivery of comprehensive preventive healthcare services across the U.S. population is not consistent, stopping both patients and healthcare providers from prioritizing regular health services (Nkouaga, 2025; Ricciardi et al., 2025).

This gap is mainly due to structural, financial, and geographic barriers that restrict access to care for millions of Americans with chronic conditions, leading to delays in early detection and treatment (Ratnapradipa et at, 2023). As a result, many Americans do not get the recommended preventive screenings and interventions, which leads to worse health outcomes, increased hospitalizations, and ongoing health disparities based on race, socio-economic status, and geography (Office of Disease Prevention and Health Promotion, n.d.). Preventable chronic conditions that could be addressed through proper screening, early detection, and behavior changes still cause most deaths and healthcare costs (Hacker, 2024).

2.2 Structural Barriers to Preventive Care Access

Among these barriers, health insurance coverage plays a key role in determining access to preventive care. Garfield et al. (2019) found that uninsured individuals are more likely than those with insurance to delay or skip treatment and are less likely to have had a preventive visit in the previous year. Recent research by Keisler-Starkey and Bunch (2024) showed that about 26 million Americans were uninsured as of 2023, with a high number in states that did not expand Medicaid under the Affordable Care Act (ACA). These differences highlight the gaps in coverage despite efforts to improve access. Blumenthal et al. (2020) showed that coverage gains under the ACA led to increased use of preventive services among newly insured

people, reinforcing that insurance is essential for accessing preventive care. However, having insurance alone does not guarantee access to preventive healthcare. The supply and distribution of primary care providers worsen this problem. The Association of American Medical Colleges (2021) predicted a shortage of physicians ranging from 17,800 to 48,000 by 2034, with the biggest shortages in rural areas and underprivileged urban communities. Studies show that this has direct health consequences, as areas with fewer primary care doctors had significantly higher death rates from heart disease, diabetes, and cancer, establishing a clear link between provider availability and preventable deaths (Basu et al., 2019).

In addition, transportation barriers continue to limit access to preventive healthcare. According to Wolfe et al. (2020), transportation issues increased sharply between 1997 and 2017, especially for low-income and rural populations. Henning-Smith et al. (2019) found that rural populations were more likely to report transportation problems as a reason for delayed care. These financial, workforce, and geographic barriers affect the same communities and groups, collectively ensuring that those who need preventive care the most are least likely to receive it.

2.3 Racial and Ethnic Disparities in Healthcare

Another challenge is racial and ethnic disparities in the detection of chronic diseases and access to preventive care. This is one of the most persistent problems in the U.S. healthcare system, driven by structural racism through housing policies, labor market inequalities, and how healthcare is funded and delivered (Bailey et al., 2021). Williams et al. (2019) reviewed evidence connecting both individual and structural racism to poor health outcomes in cardiovascular, metabolic, and mental health conditions, concluding that racism must be viewed as a fundamental cause of health disparities rather than just a factor related to socio-economic status. Yearby (2020) argued that the social determinants of health framework must be changed to identify structural racism as a root cause rather than treating race as another social variable alongside income or education. Hardeman et al. (2022) said that dismantling structural racism in healthcare requires systemic changes in how care is organized, funded, and delivered, and Snyder et al. (2023) supported this by showing that having more Black doctors in primary care was linked to lower life expectancy gaps, suggesting that the makeup of the healthcare workforce is itself a mechanism that can either help or hinder equitable access to chronic disease detection.

2.4 Predictive Analytics for Prevention and Health Equity

Tackling this challenge requires more than small policy changes. It demands innovative, data-driven tools that can identify vulnerable populations, map the structural barriers to preventive care, and provide evidence-based recommendations that policymakers can use effectively. Predictive analytics is the use of statistical modeling, machine learning, and large-scale population data to predict disease trends and risk patterns, which has emerged as one of the most promising tools for this purpose (Nwoke, 2024; Luo et al., 2016). In contrast to traditional frameworks that wait until health issues arise before taking action, predictive analytics offers a new, proactive way of dealing with health challenges. It helps identify which people are most at risk, why they are at risk, and what targeted actions can help bridge the gap between the care they receive and the care they need (Badawy et al., 2023).

Wesson et al. (2022) highlighted that the biggest benefit of using predictive analytics in public health is identifying the broader structural issues that affect entire communities. Addressing these issues can reduce

the risk of chronic diseases on a larger scale, rather than focusing only on individual patients. This perspective is supported by the growing use of predictive analytics across various parts of the U.S. public health system.

2.5 Applications for Predictive Analytics in Healthcare and Policy

The application of predictive analytics in healthcare also raises important questions concerning data governance, privacy, and regulatory accountability. Ogunsakin and Mukasa (2026) examined the ethical and regulatory challenges associated with AI-enabled health-cloud systems in the United States, identifying patient data protection, regulatory uncertainty, and stakeholder trust as important considerations for responsible implementation. Although their study focuses on cloud-based healthcare technologies rather than survey-based chronic disease analysis, it highlights broader governance challenges relevant to translating predictive insights into operational healthcare and public policy applications. These considerations reinforce the importance of ensuring that future population-health decision-support systems balance analytical capabilities with transparency, appropriate data safeguards, and institutional accountability.

At the federal level, the Centers for Disease Control and Prevention (CDC) launched the Data Modernization Initiative (DMI), a long-term project aimed at modernizing public health data systems. The initiative uses large-scale data integration, cloud computing, and predictive modeling to improve data management. One of its key applications is the National Diabetes Prevention Program Operations Center, which uses predictive analytics to track program reach, find areas that are not being covered, and give real-time risk assessment tools to organizations providing lifestyle change programs. The DMI also automates cancer reporting to provide near-real-time data, which cuts the time between collecting data and making it usable for policy decisions from six to eight weeks to just two to three days (Carney et al., 2023).

At the health system level, Rojas et al. (2022) found through a survey of large U.S. health systems that predictive analytics has become widely used, primarily for tracking hospital readmissions, detecting clinical deterioration, and assessing chronic disease risks. The Centers for Medicare and Medicaid Services (CMS) has also encouraged this use, supporting it through programs like the Medicare Shared Savings Program, which offers financial rewards to health systems that use predictive risk modeling to lower preventable hospitalizations for people with chronic conditions (CMS, n.d.).

Academic institutions have also made significant contributions to the use of predictive models that include social determinants of health (SDOH). Zhao et al. (2021) from New York University demonstrated that machine learning models incorporating factors such as income, insurance status, marital status, and neighborhood conditions are better at predicting cardiovascular disease in diverse populations than models that rely only on clinical data, and this was published in the American Journal of Preventive Medicine. Batarseh et al. (2020) from George Mason University used big data to analyze CDC data on chronic diseases and showed that population-level predictive models can identify groups of people with similar structural risk factors. They argued that such models must be part of preventive policy rather than just clinical tools. Luo et al. (2016) reviewed the use of big data in health policy and found that large, nationally representative surveys, especially those conducted by the CDC, are the best platform for examining structural factors across multiple health conditions. Their work supports the kind of cross-condition analysis needed for effective, multi-domain preventive policies.

2.6 Population-Health-Informed Medical-Device Demand Forecasting and Supply-Chain Resilience

Predictive analytics for preventive healthcare may have applications beyond identifying associations between chronic disease diagnoses and barriers to care. Delivering screening, diagnostic, and monitoring services also requires reliable access to appropriate medical devices and consumables. This creates a potential research pathway from population-health indicators, through anticipated healthcare-service utilization, to the planning of medical-device requirements. The relationship is indirect: the prevalence of a diagnosed condition alone does not establish the number of services performed or devices consumed.

Medical-device supply-chain disruptions have been examined in federal research. The U.S. Department of Health and Human Services (HHS, 2021) described challenges affecting the production and distribution of ventilators, personal protective equipment, and diagnostic testing supplies during the COVID-19 pandemic. These experiences illustrate that healthcare demand, production constraints, and distribution capacity can become misaligned, particularly during disruptions. They do not, by themselves, establish present-day shortage levels for devices used in routine chronic disease care.

Research in supply-chain management provides additional context for examining the potential application of predictive analytics to healthcare resource planning. Fowosere et al. (2025) reviewed the role of artificial intelligence in green supply-chain management, discussing its applications in demand forecasting, resource optimization, and waste reduction. Although their analysis focuses on sustainability-oriented supply chains rather than healthcare, it illustrates how predictive methods can support operational planning and resource allocation. This provides a cross-sector conceptual basis for investigating whether population-health indicators and anticipated healthcare-service utilization can inform medical-device requirements and supply-chain planning. However, translating these principles into healthcare applications would require clinically supported device-use assumptions, operational supply data, and independent validation.

The Food and Drug Administration (FDA, n.d.-a) identifies proactive monitoring, risk assessment, advanced analytics, modeling, and collaboration as elements of its medical-device supply-chain resilience work. The FDA also maintains a medical-device shortages list (FDA, n.d.-b), although inclusion on the list does not necessarily establish that patient care has been affected. Complementary HHS research on medical-product supply chains emphasizes challenges in defining and measuring resilience, criticality, and vulnerability, as well as the importance of data transparency and coordination (HHS, 2025). Together, these sources support the policy relevance of investigating methods that connect anticipated healthcare needs with supply-chain planning.

A population-health-informed approach could be developed in stages. First, survey-based indicators would describe diagnosed chronic conditions and patterns of healthcare access and utilization. Second, additional evidence on clinical workflows, eligibility, and service frequency would support estimates of relevant screening or monitoring service volumes. Third, documented device-use rates would translate service volumes into device or consumable requirements. Finally, where operational data become available, estimated requirements could be compared with inventory, procurement lead times, replenishment capacity, and disruption indicators. Each transition would require explicit assumptions and independent validation.

Diabetes offers an illustrative starting point because it is one of the chronic conditions examined in the present NHIS analysis and can involve recurring testing and monitoring services. Nevertheless, NHIS responses indicating a prior diabetes diagnosis do not reveal the clinical appropriateness, frequency, or type of device use. The study also uses an unweighted analytic sample; it therefore does not produce a nationally representative device-demand estimate. Subsequent work would need survey-design-aware estimation,

clinically supported service assumptions, and device-specific operational data before producing actionable forecasts.

The literature thus motivates a prospective integration of population-health analytics with medical-device demand and supply-chain modeling, while leaving the empirical link between these domains to be tested. The present study contributes evidence on diagnosed chronic conditions, structural access factors, and healthcare utilization.

2.7 Gaps in Literature

Taken together, these examples show that predictive analytics has moved from being a theoretical concept to a real part of U.S. healthcare. The main challenge now is using the insights from these models to create lasting policy changes that can systematically address the chronic disease gaps caused by structural barriers. Despite growing research on how social determinants of health affect chronic disease outcomes, much of this research remains descriptive and focused on single diseases or isolated issues, such as insurance, transportation, or racial disparities. While studies on predictive analytics and AI in healthcare show the potential of data tools for identifying risk, many of them focus on clinical decision-making, model accuracy, or how health systems operate, rather than on policy applications.

Therefore, there is still a need to understand how predictive analytics can be used with nationally representative U.S. data to measure how social determinants of health influence chronic conditions and how these findings can lead to better preventive policies. A further opportunity for future research is to investigate whether population-health and preventive-care utilization indicators can improve healthcare-service and device-demand planning when combined with clinical workflows and supply-chain information.

3. Research Question and Objectives

The question this study aims to answer is: How can predictive analytics be used to inform and shape public policy in order to improve access to preventive healthcare services for chronic conditions in the United States?

To answer this question, the study pursues four objectives:

1. Determine how demographic, socioeconomic, and healthcare access factors relate to time since the last doctor visit and wellness-check attendance.
2. Determine which of these factors are associated with five diagnosed conditions (hypertension, diabetes, cardiovascular disease, high cholesterol, and cancer), with and without healthcare utilization measures in the models.
3. Compare each model's classification accuracy on held-out test data against a majority-class baseline, cross-checked with Random Forest classifiers.
4. Derive differentiated recommendations for insurance policy, screening infrastructure, and risk tools built on structural data.

4. Methods

Data Source

This study uses data from the 2024 National Health Interview Survey (NHIS) Sample Adult File (N = 32,629), a nationally representative, in-person household survey administered by the National Center for Health Statistics (NCHS). The survey year was confirmed directly from the dataset's SRVY_YR field. The NHIS uses a multistage probability sampling design with oversampling of Black, Hispanic, and Asian adults; full design and estimation procedures are documented in Parsons et al. (2014).

Variable Selection and Operationalization

Sixteen variables were selected from the full NHIS Sample Adult File and organized into four conceptual categories consistent with the study's guiding framework:

Demographic control variables: age (AGEP_A), sex (SEX_A), race (RACEALLP_A), Hispanic ethnicity (HISP_A), educational attainment (EDUCP_A), marital status (MARITAL_A), and census region (REGION).

Healthcare access independent variables: health insurance coverage (HICOV_A), usual source of care (USUALPL_A), and reliable transportation (TRANSPOR_A).

Chronic condition dependent variables: hypertension (HYPEV_A), diabetes (DIBEV_A), cardiovascular disease (CHDEV_A), high cholesterol (CHLEV_A), and cancer (CANEV_A), each representing self-reported prior diagnosis ("ever told by a health professional").

Healthcare utilization variables: time since last doctor visit (LASTDR_A) and whether the most recent visit was specifically a wellness/routine checkup (WELLNESS_A). These two variables were treated as dependent variables in an initial analysis stage, and subsequently incorporated as additional independent variables in a later modeling stage (described below).

All variables were renamed for interpretability and recoded from their raw NHIS numeric codes into labeled factors. Reference categories were deliberately selected to support intuitive interpretation: "No" was set as the reference level for all binary condition and access variables (so that coefficients represent the odds of the condition being present or the barrier being encountered), and "Within the past year" was set as the reference level for time since last visit, both because it was the modal category and to avoid the estimation instability associated with rarer reference categories (detailed below).

Missing Data

NHIS employs a standard missing-data coding convention in which single-digit variables use codes 7 (Refused), 8 (Not Ascertained), and 9 (Don't Know), while two-digit variables use 97, 98, and 99. These codes were identified and re-coded to NA for every study variable prior to analysis. Age (AGEP_A), while topcoded at 85 ("85 and older"), was confirmed via inspection of its value distribution to also follow the two-digit missing-code convention (97, 98, 99), rather than containing genuine ages in that range. The inspection showed a complete absence of any recorded ages between 86 and 96, confirming no genuine ages were recorded in that range.

Following missing-value re-coding, listwise deletion (complete-case analysis) was applied, retaining only respondents with non-missing values across all sixteen study variables. This reduced the analytic sample from 32,629 to 29,194 respondents (89.5% retention). When the wellness-check variable was later incorporated, the analytic sample was further reduced to 29,110, reflecting the variable's structural missingness among respondents who reported never having seen a doctor (for whom no "last visit" characteristics exist).

Survey Weights

All analyses were conducted on the unweighted complete-case sample. Survey design weights and stratum/cluster identifiers were not applied, because the study's objectives concern the estimation of conditional associations and the evaluation of out-of-sample classification accuracy rather than the production of nationally representative prevalence estimates, and because the train/test partitioning used for predictive evaluation does not accommodate the NHIS design-based variance estimation framework. As a consequence, the descriptive statistics and association estimates reported here should be interpreted as applying to the analytic sample rather than as design-based national estimates.

Category Consolidation

The NHIS time-since-last-visit variable (LASTDR_A) contains seven ordered categories, including a "Never" category for respondents who report never having seen a doctor. Initial multinomial modeling revealed that this category could not be estimated reliably: it comprised only 46–53 respondents (0.14–0.18% of the analytic sample), and for several demographic subgroups, such as American Indian/Alaska Native respondents and respondents with an occupational associate degree, it contained zero observations. Because some predictor levels therefore perfectly predicted category membership, maximum likelihood estimation had no stable interior solution, producing non-estimable coefficients with odds ratios of 0 or values exceeding 10^7 and degenerate confidence intervals, the signature of complete or quasi-complete separation. The problem was structural rather than sample-size-dependent, since refitting the model on the full complete-case sample and on a 75% training subsample reproduced the same pattern in both. The "Never" category was therefore merged with the adjacent category "10+ years ago or more," as the two occupy the same position on the underlying care-engagement continuum, representing long-term or complete disengagement from routine care. The resulting six-category outcome was fully estimable and stable and was used in all subsequent LASTDOCVISIT models.

Modeling Approach

The four analytical stages described below operationalize the study's objectives (Section 3): Stage 1 corresponds to Objective 1 (access barriers to utilization), Stages 2 and 3 to Objective 2 (structural predictors of diagnosis), and Stage 4, together with the train/test evaluation, to Objective 3 (predictive accuracy against naive benchmarks); Objective 4 is addressed in the Policy Implications.

Analysis proceeded in four stages:

Stage 1 (Utilization models): Two models assessed which demographic and access factors predict healthcare utilization: (a) a multinomial logistic regression of time since last doctor visit (6 categories) on the ten demographic and access variables, and (b) a binary logistic regression of wellness-check status on the same predictor set.

Stage 2 (Baseline chronic condition models): Five binary logistic regression models assessed the association between the same ten demographic/access predictors and each chronic condition (hypertension, diabetes, cardiovascular disease, cholesterol, cancer).

Stage 3 (Extended chronic condition models): Each of the five Stage 2 models was re-estimated in three additional configurations, incorporating (a) time since last visit, (b) wellness-check status, and (c) both variables simultaneously, as additional independent variables

All models were fit using maximum likelihood estimation (glm() with a binomial link for binary outcomes; multinom() for the multinomial outcome). Results are reported as odds ratios with 95% confidence intervals, obtained by exponentiating model coefficients.

Statistical Model

Binary Logistic Regression (Chronic Conditions and Wellness Check). For each binary outcome, the probability of a positive outcome (*p*) was modeled as:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{10} X_{10}$$

where *p* is the probability of the outcome of interest (e.g., probability of hypertension, or probability the last visit was a wellness check), β_0 is the intercept, and X_1 through X_{10} are the predictors defined in Table M1.

Multinomial Logistic Regression (Time Since Last Doctor Visit). For the six-category LASTDOCVISIT outcome, a separate equation was estimated for each non-reference category *j*, relative to the reference category ("within the past year"):

$$\ln\left(\frac{p_j}{p_{ref}}\right) = \beta_{0,j} + \beta_{1,j} X_1 + \beta_{2,j} X_2 + \cdots + \beta_{10,j} X_{10}, \quad j = 1, \ldots, 5$$

Extended Binary Models (Stage 3). For the five chronic condition models re-estimated with utilization predictors added, two additional terms were included:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12}$$

Table M1. Predictor Definitions

International Journal of Computer Applications Technology and Research
Volume 14–Issue 11, 94 – 117, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1411.1008

In all models, exponentiating each coefficient (e^{β}) yields the corresponding odds ratio reported throughout

Symbol	Variable
X_1	Age
X_2	Sex
X_3	Race
X_4	Ethnicity
X_5	Region
X_6	Educational attainment
X_7	Marital status
X_8	Health insurance coverage
X_9	Usual source of care
X_{10}	Reliable transportation
X_{11}	Time since last doctor visit *(Stage 3 extended models only)*
X_{12}	Wellness-check status *(Stage 3 extended models only)*

In all models, exponentiating each coefficient (e^{β}) yields the corresponding odds ratio reported throughout this study, and predicted probabilities were obtained by inverting the logit via the sigmoid function:

$$p = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i X_i)}}$$

Classification for binary outcomes was performed by applying a 0.5 threshold to this predicted probability; for the multinomial outcome, the predicted category was the one with the highest estimated probability across all six categories.

Train/Test Evaluation

To distinguish statistical association from genuine predictive utility, the complete-case sample was randomly partitioned into a 75% training set (n = 21,895–21,964, depending on model) and a 25% held-out test set (n = 7,278–7,344), using a fixed random seed for reproducibility. All models were fit exclusively on the training partition. Predictive performance was then assessed on the test partition using classification accuracy (at a 0.5 predicted-probability threshold) and compared against a majority-class baseline. This baseline comparison was essential given substantial prevalence imbalance across outcomes (ranging from

approximately 6.5% for cardiovascular disease to 87% for the modal LASTDOCVISIT category), under which raw accuracy alone can substantially overstate genuine model performance.

Stage 4: Random Forest Machine Learning Classifiers

To complement and validate the logistic regression findings, Random Forest classifiers (Breiman, 2001) were trained on the same five chronic condition outcomes, using the same 75/25 train/test split and identical predictor sets. Random Forest is an ensemble method that builds B = 500 decision trees on bootstrap samples of the training data and aggregates their predictions by majority vote; the number of predictors randomly sampled at each split was set to mtry = 3, approximately the square root of the total predictor count. Unlike logistic regression, Random Forest does not assume a linear relationship between predictors and outcomes. The primary purpose of including it is to independently validate the structural predictor rankings identified in logistic regression through a fundamentally different analytical mechanism, and to assess whether a more flexible model can improve predictive accuracy for the rare-prevalence conditions where logistic regression was constrained by class imbalance. For the cancer models, class weighting (classwt = No:1, Yes:5) was applied to address severe class imbalance, penalizing misclassification of cancer cases five times more heavily than non-cases.

Variable importance was assessed using Mean Decrease Accuracy (MDA): for each predictor, its values were randomly permuted across each tree's out-of-bag observations, and the resulting average reduction in out-of-bag accuracy across all 500 trees was recorded as that predictor's importance score. Higher MDA values indicate greater variable importance; negative values indicate that permuting the variable actually improves accuracy. All Random Forest models were estimated using the randomForest package in R.

$$Gini(t) = 1 - [p(Yes \mid t)^2 + p(No \mid t)^2]$$

where $p(Yes \mid t)$ and $p(No \mid t)$ are the proportions of Yes and No cases at node t. The prediction for a new observation x was determined by majority vote across all 500 trees, and predicted class probabilities were estimated as the proportion of trees voting for each class. For the cancer models, class weighting (classwt = No:1, Yes:5) was applied to address severe class imbalance, penalizing misclassification of cancer cases five times more heavily than non-cases.

Variable importance was assessed using Mean Decrease Accuracy (MDA), defined as the average reduction in out-of-bag accuracy when a predictor is randomly permuted across all 500 trees:

$$MDA(X_j) = \frac{1}{1500}\sum_{b}[Accuracy(T_b, OOB_b) - Accuracy(T_b, OOB_b \cdot permuted\ \text{j})$$

where OOB_b denotes the out-of-bag observations not used in bootstrap sample b.

5 Results and Discussion

The analysis proceeded in four stages, and the results are presented in that order; a synthesis follows.

Stage 1 (utilization). In the multinomial model of time since last doctor visit, usual source of care was the dominant predictor: respondents with no usual source of care showed sharply reduced odds of recent visits, an effect strengthening with time since visit (odds ratios from approximately 0.17 at "within the last 2 years" to 0.03 at "10+ years ago"), with insurance coverage following a similar but smaller pattern. In the wellness-check model, usual source of care (OR ≈ 0.47), insurance coverage (OR ≈ 0.70), and reliable transportation (OR ≈ 1.32; direction discussed in Stage 2) were significant, while age, sex, and education were not, suggesting that wellness-check attendance is governed more by structural access conditions than by demographic characteristics. Notably, neither utilization model exceeded its majority-class baseline (LASTDOCVISIT: 88.06% vs. 88.10%; wellness check: 82.47% vs. 82.47%, with the fitted model predicting "Yes" for the entire test sample). This first demonstration of the significance–prediction distinction establishes the methodological theme that recurs below.

Stage 2 (baseline condition models): Age was significantly positively associated with every condition. Usual source of care was the most consistent access predictor across all five models (ORs 0.34–0.61 for "no usual place"), interpreted throughout as a **detection effect**: because all outcomes measure diagnosed disease, reduced healthcare contact mechanically suppresses diagnosis independent of underlying prevalence. Racial and educational disparities were condition-specific: Black respondents showed elevated odds of hypertension (OR = 1.67) and diabetes (OR = 1.58) but reduced odds of cancer diagnosis (OR = 0.58), a reversal also observed for Asian and Hispanic respondents; higher education was protective for hypertension, diabetes, cardiovascular disease, and cholesterol (ORs as low as 0.39–0.46) but associated with elevated cancer odds (ORs up to 1.82). Transportation displayed the opposite reversal: lacking reliable transportation predicted higher diagnosis odds for hypertension, diabetes, cardiovascular disease, and cholesterol (ORs 1.18–1.65), while lacking insurance or a usual source of care predicted lower odds. Classification accuracy exceeded the majority-class baseline only for the two most prevalence-balanced conditions, hypertension (72.0% vs. 60.7%, +11.3 points) and cholesterol (68.8% vs. 65.5%, +3.4 points); the rarer conditions (diabetes 11%, cancer 14%, cardiovascular disease 6.5%) performed indistinguishably from baseline despite numerous individually significant predictors.

Stage 3 (extended models): Time since last visit was significant in essentially every category of all five condition models, displacing the significance of insurance coverage for hypertension and cardiovascular disease and of usual source of care for cancer; wellness-check status was significant only for cardiovascular disease (OR ≈ 0.69–0.70) and cancer (OR ≈ 0.66–0.67), both $p < .001$. Yet the Stage 2 accuracy pattern held with near-perfect consistency (Table 1): the balanced-prevalence conditions showed modest compounding gains (up to +3.97 points for the fully specified cholesterol model), while the rare conditions showed essentially zero change across every specification, including cardiovascular disease's "+Both" model, in which two independently significant predictors (both $p < .001$) were added simultaneously and accuracy nonetheless remained at or below baseline. This is the strongest evidence in the study that the limitation is a structural property of outcome prevalence rather than an artifact of an insufficiently rich predictor set: a repeatedly confirmed, clinically plausible predictor could not overcome the classification ceiling imposed by a 6.5% base rate at a 0.5 threshold.

Table 1: Classification accuracy across all four model versions.

International Journal of Computer Applications Technology and Research
Volume 14–Issue 11, 94 – 117, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1411.1008

Condition	Prevalence	Baseline Model	+LASTDOCVISIT	+WellnessCheck	Both LASTDOCVISIT and WellnessCheck
Hypertension	~39%	72.03% (+11.3)	72.60% (+11.9)	72.01% (+11.3)	72.60% (+11.9)
Cholesterol	~33%	68.82% (+3.4)	69.35% (+3.9)	68.77% (+3.3)	69.43% (+4.0)
Cancer	~14%	85.50% (0.0)	85.49% (0.0)	85.64% (+0.2)	85.60% (+0.1)
Diabetes	~11%	88.66% (0.0)	88.66% (0.0)	88.66% (0.0)	88.66% (0.0)
Cardiovascular Disease	~6.5%	93.29% (0.0)	93.29% (0.0)	93.27% (0.0)	93.27% (0.0)

(Values in parentheses indicate percentage-point improvement over the majority-class baseline.)

The utilization predictors also revealed an attenuation pattern: insurance coverage, significant at baseline for hypertension and cardiovascular disease ($p = .009$ and $p = .025$), lost significance once visit recency was included ($p = .790$ and $p = .108$), with a parallel pattern for usual source of care in the cancer model ($p < .001$ to $p = .107$). This is consistent with a mediating role of utilization, in which structural access operates on diagnosis substantially through visit frequency; rigorously establishing such a pathway would require formal mediation analysis and longitudinal data. The pattern was not universal: insurance remained significant for cholesterol throughout ($p < .001$), suggesting cholesterol screening is less tightly coupled to general visit recency. For policy, interventions at different points along this pathway may yield overlapping rather than additive benefits.

Stage 4 (Random Forest cross-validation): Random Forest accuracy was fully consistent with the regression results (Table 2). Hypertension (0.7292) and cholesterol (0.6979) exceeded both baseline and logistic regression; diabetes (0.8866) and cardiovascular disease (0.9328) matched all three benchmarks exactly, confirming that the class imbalance ceiling cannot be overcome by a more flexible model using structural variables alone; and cancer accuracy fell to 0.5982 because class weighting forced the model to attempt minority-class detection at the cost of aggregate accuracy, the appropriate tradeoff where detecting true cases matters more than maximizing raw accuracy.

Recent research on healthcare risk forecasting has highlighted the potential of machine-learning approaches to capture nonlinear relationships and complex risk patterns that may not be adequately represented by conventional statistical models. Mukasa (2026), in a review of AI-driven Medicare risk-score forecasting, discusses these methodological opportunities while emphasizing the need for appropriate evaluation and transparency. Although Medicare risk adjustment differs from chronic disease classification, this perspective provides relevant context for comparing alternative predictive methods. Accordingly, the present study evaluates Random Forest classifiers alongside logistic regression using the same chronic disease outcomes, predictor sets, and held-out test partitions. Performance is assessed against majority-class baselines to determine whether greater modeling flexibility produces meaningful improvements in classification accuracy.

Table 2: Model Accuracy Comparison

Condition	Baseline Acc.	LR Accuracy	RF Accuracy	Interpretation
Hypertension	0.6069	0.726	0.7292	Both models beat baseline; RF slightly exceeds LR
Diabetes	0.8866	0.8866	0.8866	All identical; class imbalance ceiling
Cardiovascular Disease	0.9328	0.9327	0.9328	All essentially identical; severe class imbalance
High Cholesterol	0.6546	0.6943	0.6979	Both models beat baseline; RF slightly exceeds LR
Cancer	0.8549	0.856	0.5982	RF drops below baseline; class weighting forces minority class detection

(LR = Logistic Regression; RF = Random Forest. Baseline accuracy reflects the proportion of the test set belonging to the majority class.)

International Journal of Computer Applications Technology and Research
Volume 14–Issue 11, 94 – 117, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1411.1008

The variable importance plots (Figures 1–5) independently confirmed the logistic regression predictor structure through a completely different mechanism. Age dominated every condition (largest for hypertension, approximately 260 MDA points, and cholesterol, approximately 175 points). Visit recency and usual source of care ranked second and third for hypertension and cholesterol, the clearest structural access pattern of the five conditions, consistent with the dependence of cholesterol detection on a clinical encounter. Race and Hispanic ethnicity ranked above all access variables for diabetes and cardiovascular disease; for the latter, the transportation barrier ranked above insurance, consistent with the Stage 2 transportation reversal. For cancer, usual source of care and insurance showed negative importance, meaning that removing them improved accuracy: a direct independent confirmation of the cancer detection reversal, indicating that cancer detection depends on specific screening encounters rather than routine primary care contact. Wellness-check status was the only access predictor with positive importance for cancer, consistent with Stage 3.

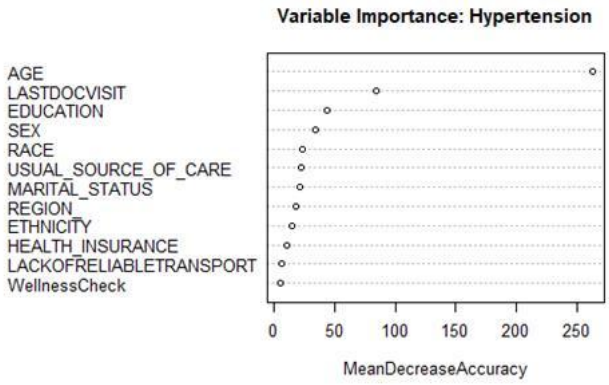
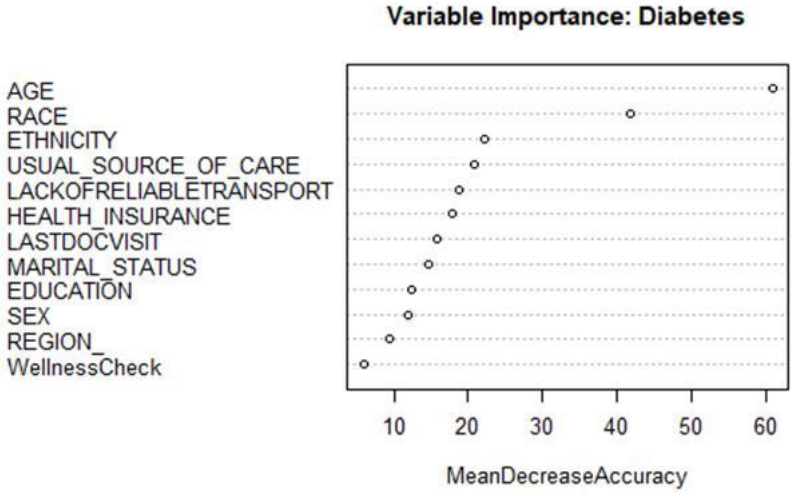


Figure 1: Random Forest variable importance for hypertension (Mean Decrease Accuracy). Age dominates at approximately 260 points. Last doctor visit and education are the leading structural predictors, confirming logistic regression findings.



International Journal of Computer Applications Technology and Research
Volume 14–Issue 11, 94 – 117, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1411.1008

Figure 2: Random Forest variable importance for diabetes (Mean Decrease Accuracy). Age dominates. Race ranks second and Hispanic ethnicity third , both above all healthcare access variables , confirming racial and ethnic disparities in diabetes detection.

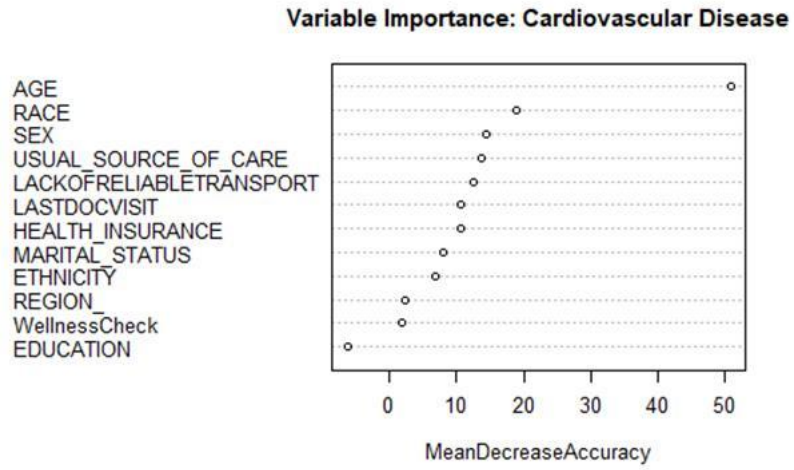


Figure 3: Random Forest variable importance for cardiovascular disease (Mean Decrease Accuracy). Age dominates. Race ranks second above all access variables. Transportation barrier outranks health insurance.

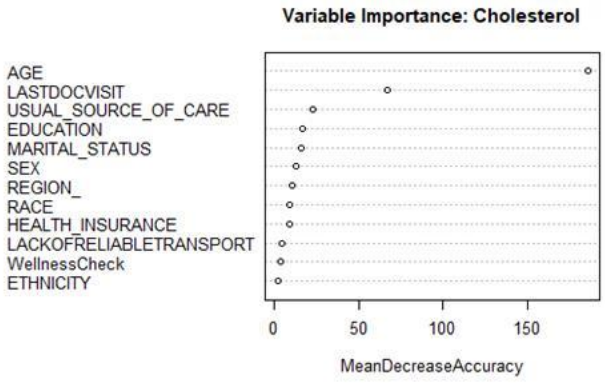


Figure 4: Random Forest variable importance for high cholesterol (Mean Decrease Accuracy). Age dominates. Last doctor visit and usual source of care rank second and third, the clearest structural access pattern of all five conditions.

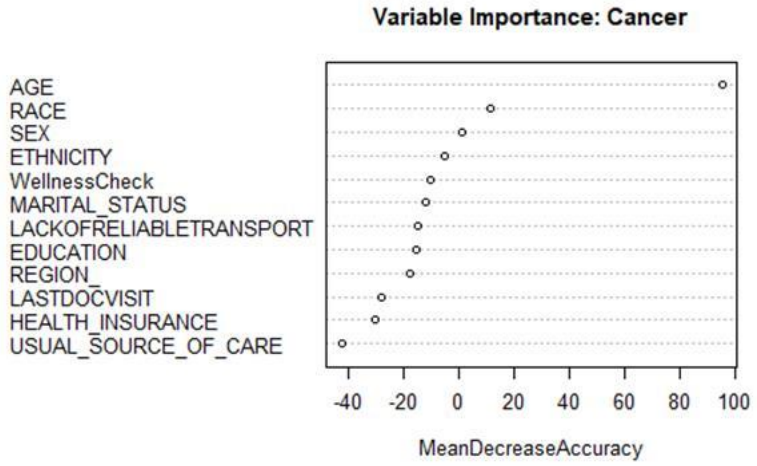


Figure 5: Random Forest variable importance for cancer (Mean Decrease Accuracy). Age dominates. Usual source of care and health insurance show negative importance, independently confirming the cancer detection reversal pattern identified in logistic regression.

Synthesis: Across 22 regression specifications and five Random Forest models, demographic and healthcare access variables were associated with utilization and reported chronic disease diagnoses. The largest classification accuracy improvements over majority-class baselines occurred for hypertension and high cholesterol, while improvements for the less prevalent conditions were absent or small. These findings can inform preventive-care priorities within the limits of the study design, but broader performance evaluation and further validation are needed before applying the models to individual risk assessment. Further research should test whether additional clinical or biometric information improves performance rather than assuming that such information is necessarily required.

6. Policy Recommendation

- **Strengthen Continuity of Care Alongside Insurance Expansion:** Policymakers should pair insurance expansion with measures that help individuals establish and maintain a usual source of care. Patient-centered medical homes and community health worker programs offer approaches for strengthening connections to preventive services, particularly among populations facing access barriers (John et al., 2020; Berini et al., 2022).

- **Address Potential Screening Gaps Among Underserved Populations:** Healthcare systems should investigate whether lower recorded diagnosis rates among uninsured individuals and those without a usual source of care reflect unmet screening needs. Targeted outreach, assistance connecting patients with regular care, and mobile screening where appropriate could improve access. These initiatives should be evaluated through screening uptake, diagnostic follow-up, and newly identified diagnoses.

- **Develop Condition-Specific Cancer Screening Strategies:** Policymakers should assess barriers to recommended cancer screening separately from broader chronic disease prevention efforts. The distinct racial and educational patterns observed in this study support further investigation of differences in screening and diagnostic access. Outreach and patient navigation should follow established screening eligibility criteria and address barriers to completing follow-

up care. Broader research on structural racism provides context for examining these institutional barriers (Bailey et al., 2021; Yearby, 2020).

- **Strengthen Evaluation of Predictive Risk Tools:** Healthcare systems should evaluate predictive models using measures beyond overall accuracy, including sensitivity, specificity, precision, discrimination, and calibration. Evaluation should account for outcome prevalence and differences in performance across population groups. Further research should test whether combining demographic and access variables with clinical or biometric information improves individual-level prediction, building on research examining social determinants in cardiovascular risk models (Zhao et al., 2021).

- **Conduct Further Research on Disease Detection and Policy Effectiveness:** Future studies should use longitudinal data and linked clinical records to examine whether access barriers contribute to delayed or missed diagnoses. Evaluations should also assess whether stronger connections to primary care and targeted preventive outreach improve screening, follow-up, and health outcomes. This would help determine which policy approaches produce sustained improvements in preventive healthcare access.

7. Population-Health-Informed Medical-Device Demand Forecasting and Supply-Chain Resilience

7.1 Rationale and relationship to the present findings

The present analysis identifies associations among demographic characteristics, access to care, healthcare utilization, and self-reported chronic disease diagnoses. These findings suggest a potential upstream input for healthcare resource planning: estimates of the populations likely to seek or require particular services. Such an application would not equate diagnosed disease with true underlying prevalence. The observed access-related differences in reported diagnoses could reflect differential detection, and the diabetes classifiers did not improve on the majority-class accuracy baseline.

7.2 Conceptual architecture

The proposed framework connects five analytically distinct stages: (1) characterize population-health indicators and access patterns; (2) estimate disease-relevant healthcare-service volumes; (3) translate projected services into requirements for specified medical devices or consumables using documented clinical workflows; (4) compare requirements with inventory, replenishment lead times, procurement, and supply-risk indicators when available; and (5) generate interpretable planning scenarios and potential early-warning signals. Each stage introduces additional data requirements and uncertainty, which must be recorded and evaluated separately.

The empirical evidence in this paper is limited to the 2024 NHIS analytic sample and its fitted association and classification models. Implementing the extension would require additional sources for eligible-population estimates, clinical practice and device-use parameters, utilization trends, device inventories, procurement lead times, and potentially manufacturer or distributor capacity. Public FDA medical-device supply-chain materials establish the policy relevance of resilience monitoring, but do not substitute for product-level operational data. This framework is therefore presented as a research and implementation agenda, not a validated medical-device forecasting system.

NHIS population-health and access indicators → Service-volume scenarios → Device-use assumptions and requirements → Supply availability and risk comparison → Planning and policy decision support

International Journal of Computer Applications Technology and Research
Volume 14–Issue 11, 94 – 117, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1411.1008

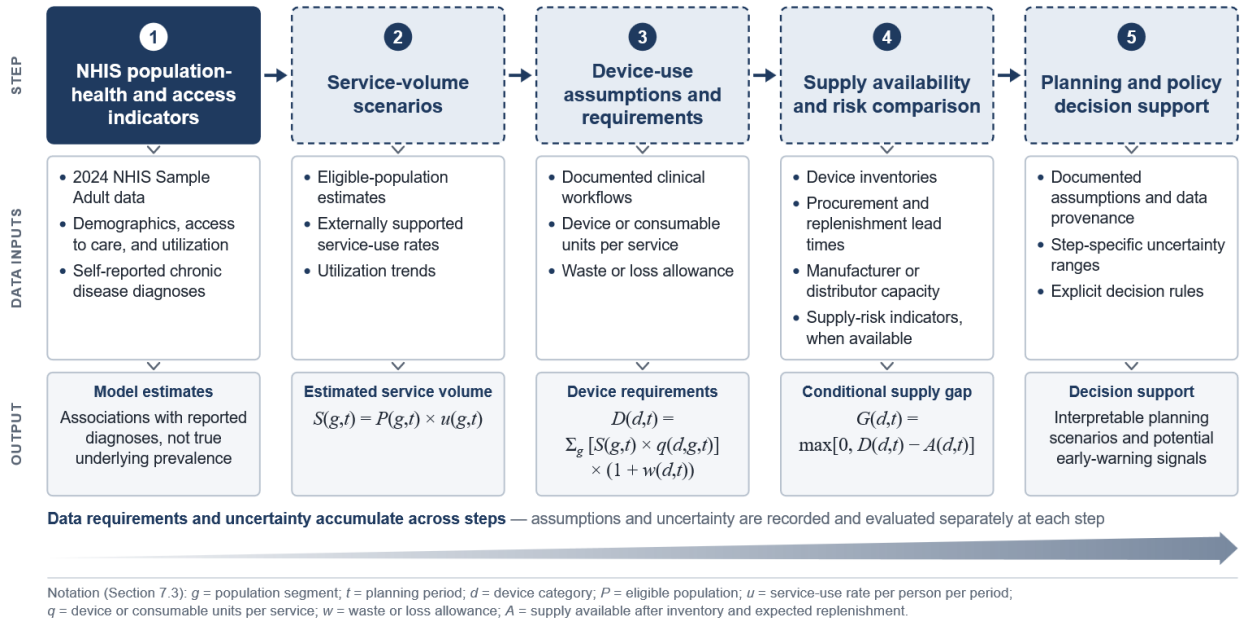


Figure 6: Proposed framework linking population-health indicators to medical-device demand and supply-chain resilience planning.

7.3 Illustrative model and assumptions

For population segment g and planning period t, estimated service volume can be defined as S(g,t) = P(g,t) × u(g,t), where P is the appropriately estimated eligible population and u is an externally supported service-use rate per person per period. For device category d, estimated requirements can be defined as D(d,t) = Σg [S(g,t) × q(d,g,t)] × (1 + w(d,t)), where q is device or consumable units per service and w is a documented allowance for waste or other losses. A conditional supply gap may be defined as G(d,t) = max[0, D(d,t) − A(d,t)], where A represents supply available over the same period after accounting for inventory and expected replenishment.

7.4 Validation, transferability, and policy relevance

The framework should first be tested for internal consistency and sensitivity to uncertain inputs, then evaluated against observed service volumes, purchasing records, and stock availability if partner data become accessible. Relevant metrics may include absolute and percentage forecast error, calibration of uncertainty ranges, subgroup robustness, and lead time or precision of warnings where verified supply-gap events exist. Reproducible variable definitions, assumptions, data provenance, and decision rules would allow the methodology to be assessed for transfer to other chronic conditions or device categories. Its prospective policy value lies in linking anticipated healthcare-service needs with supply planning; the present paper does not demonstrate shortage prevention, procurement savings, or improved patient outcomes.

8. Limitations

Four limitations bound these findings. First, the cross-sectional, self-reported design precludes causal claims: outcomes reflect diagnosed rather than true disease, and the detection-effect interpretation, while strongly motivated by the access literature, cannot distinguish detection from prevalence directly. Second, the attenuation findings are consistent with mediation by visit recency but establishing that pathway

rigorously would require formal mediation analysis and longitudinal data. Third, analyses were conducted on the unweighted complete-case analytic sample, so estimates describe the sample rather than the design-based national population. Fourth, the structural survey variables available in the NHIS exclude clinical and biometric measures, so the class-imbalance accuracy ceiling observed for rarer conditions reflects the predictor set available to a household survey, not an absolute limit on prediction from richer data.

9.Conclusion

This study examined how predictive analytics can inform public policy to improve preventive healthcare access for chronic conditions in the United States. Across 27 fitted models using the 2024 National Health Interview Survey Sample Adult dataset, logistic regression and Random Forest classifiers identified demographic and healthcare access factors associated with utilization and reported chronic disease diagnoses. Usual source of care and visit recency emerged as important factors, supporting policy attention to continuity of care alongside insurance coverage. However, classification accuracy exceeded majority-class baselines only for hypertension and high cholesterol under the evaluation settings used. For diabetes, coronary heart disease, and cancer, adding utilization variables or using Random Forest did not produce accuracy gains over those baselines. These findings highlight the distinction between identifying statistical associations and demonstrating useful individual-level prediction. The cross-sectional, self-reported data also limit causal interpretation and cannot directly identify undiagnosed disease. Overall, the study supports using population survey analysis to inform preventive-care priorities, while emphasizing the need for broader performance evaluation and further validation before applying these models to individual risk assessment. As a prospective extension, a population-health-informed medical-device planning framework could connect appropriately estimated service needs with device-use assumptions and independently collected supply-chain data. Its utility, accuracy, and broader applicability remain to be established through additional datasets and external validation.

References

Acquah, I., Hagan, K., Valero-Elizondo, J., Javed, Z., Butt, S. A., Mahajan, S., Taha, M. B., Hyder, A. A., Mossialos, E., Cainzos-Achirica, M., & Nasir, K. (2022). Delayed medical care due to transportation barriers among adults with atherosclerotic cardiovascular disease. *American Heart Journal, 245*, 60–69. https://doi.org/10.1016/j.ahj.2021.11.019

Andersen, R. M. (1995). Revisiting the behavioral model and access to medical care: Does it matter? *Journal of Health and Social Behavior, 36*(1), 1–10. https://doi.org/10.2307/2137284

Association of American Medical Colleges. (2021, June 11). *AAMC report reinforces mounting physician shortage*. https://www.aamc.org/news/press-releases/aamc-report-reinforces-mounting-physician-shortage

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*(3), 411–421. https://doi.org/10.1097/01.psy.0000127692.23278.a9

Badawy, M., Ramadan, N., & Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: A survey. *Journal of Electrical Systems and Information Technology, 10*, Article 40. https://doi.org/10.1186/s43067-023-00108-y

Bailey, Z. D., Feldman, J. M., & Bassett, M. T. (2021). How structural racism works—Racist policies as a root cause of U.S. racial health inequities. *New England Journal of Medicine, 384*(8), 768–773. https://doi.org/10.1056/NEJMms2025396

Basu, S., Berkowitz, S. A., Phillips, R. L., Bitton, A., Landon, B. E., & Phillips, R. S. (2019). Association of primary care physician supply with population mortality in the United States, 2005–2015. *JAMA Internal Medicine, 179*(4), 506–514. https://doi.org/10.1001/jamainternmed.2018.7624

Batarseh, F. A., Ghassib, I., Chong, D. S., & Su, P.-H. (2020). Preventive healthcare policies in the US: Solutions for disease management using Big Data Analytics. *Journal of Big Data, 7*, Article 38. https://doi.org/10.1186/s40537-020-00315-8

Benavidez, G. A., Zahnd, W. E., Hung, P., & Eberth, J. M. (2024). Chronic disease prevalence in the US: Sociodemographic and geographic variations by ZIP Code Tabulation Area. *Preventing Chronic Disease, 21*, Article E14. https://doi.org/10.5888/pcd21.230267

Berini, C. R., Bonilha, H. S., & Simpson, A. N. (2022). Impact of community health workers on access to care for rural populations in the United States: A systematic review. *Journal of Community Health, 47*(3), 539–553. https://doi.org/10.1007/s10900-021-01052-6

Bloom, D. E., Cafiero, E. T., Jané-Llopis, E., Abrahams-Gessel, S., Bloom, L. R., Fathima, S., Feigl, A. B., Gaziano, T., Mowafi, M., Pandya, A., Prettner, K., Rosenberg, L., Seligman, B., Stein, A., & Weinstein, C. (2011). *The global economic burden of non-communicable diseases*. World Economic Forum. https://www3.weforum.org/docs/WEF_Harvard_HE_GlobalEconomicBurdenNonCommunicableDiseases_2011.pdf

Blumenthal, D., Collins, S. R., & Fowler, E. J. (2020). The Affordable Care Act at 10 years—Its coverage and access provisions. *New England Journal of Medicine, 382*(10), 963–969. https://doi.org/10.1056/NEJMhpr1916091

Blumenthal, D., Gumas, E. D., Shah, A., Gunja, M. Z., & Williams, R. D., II. (2024). *Mirror, mirror 2024: A portrait of the failing U.S. health system—Comparing performance in 10 nations*. Commonwealth Fund. https://doi.org/10.26099/ta0g-zp66

Boersma, P., Black, L. I., & Ward, B. W. (2020). Prevalence of multiple chronic conditions among US adults, 2018. *Preventing Chronic Disease, 17*, E106. https://doi.org/10.5888/pcd17.200130

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Carney, T. J., Wiltz, J. L., Davis, K., Briss, P. A., & Hacker, K. (2023). Advancing chronic disease practice through the CDC Data Modernization Initiative. *Preventing Chronic Disease, 20*, E110. https://doi.org/10.5888/pcd20.230120

Centers for Disease Control and Prevention. (2023). *Chronic disease fast facts*. U.S. Department of Health and Human Services. https://www.cdc.gov/chronic-disease/data-research/facts-stats/index.html

Centers for Medicare and Medicaid Services. (2023). *Medicare Shared Savings Program: Quality and performance standards*. CMS. https://www.cms.gov/medicare/payment/fee-for-service-providers/shared-savings-program-acos

Dahima, P. (2025, July 1). *Primer: Chronic disease among adults in the United States*. American Action Forum. https://www.americanactionforum.org/insight/primer-chronic-disease-among-adults-in-the-united-states/

Dritsakis, G., Gallos, I., Psomiadi, M. E., Amditis, A., & Dionysiou, D. (2024). Data analytics to support policy making for noncommunicable diseases: Scoping review. *Online Journal of Public Health Informatics, 16*, e59906. https://doi.org/10.2196/59906

Fowosere, S., Esechie, C. O., Namboozo, S., & Anwansedo, F. (2025). The role of artificial intelligence in green supply chain management. International Journal of Latest Technology in Engineering, Management & Applied Science, 14(2), 325–331. https://doi.org/10.51583/IJLTEMAS.2025.14020033

Garfield, R., Orgera, K., & Damico, A. (2019). *The uninsured and the ACA: A primer*. Kaiser Family Foundation. https://www.kff.org/unin-sured/report/the-uninsured-and-the-aca-a-primer-key-facts-about-health-insurance-and-the-uninsured-amidst-changes-to-the-affordable-care-act/

Hacker, K. (2024). The burden of chronic disease. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes, 8*(1), 112–119. https://doi.org/10.1016/j.mayocpiqo.2023.08.005

Hardeman, R. R., Homan, P. A., Chantarat, T., Davis, B. A., & Brown, T. H. (2022). Improving the measurement of structural racism to achieve antiracist health policy. *Health Affairs, 41*(2), 179–186. https://doi.org/10.1377/hlthaff.2021.01489

Healthy People 2030. (n.d.). *Social determinants of health*. U.S. Department of Health and Human Services, Office of Disease Prevention and Health Promotion. https://health.gov/healthypeople/priority-areas/social-determinants-health

Henning-Smith C, Evenson A, Corbett A, Kozhimannil K, Moscovice I. (2017). Rural transportation: challenges and opportunities. University of Minnesota Rural Health Research Center. November 2017. Available at: https://rhrc.umn.edu/wpcontent/uploads/2019/01/1518734252UMRHRCTransportationChallenges.pdf

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387

John, J. R., Jani, H., Peters, K., Agho, K., & Tannous, W. K. (2020). The effectiveness of patient-centred medical home-based models of care versus standard primary care in chronic disease management: A systematic review and meta-analysis of randomised and non-randomised controlled trials. *International Journal of Environmental Research and Public Health, 17*(18), 6886. https://doi.org/10.3390/ijerph17186886

Kaiser Family Foundation. (2024). *Key data on health and health care by race and ethnicity*. https://www.kff.org/racial-equity-and-health-policy/key-data-on-health-and-health-care-by-race-and-ethnicity/

Keisler-Starkey, K., & Bunch, L. N. (2024). *Health insurance coverage in the United States: 2023* (Current Population Reports, P60-284). U.S. Census Bureau. https://www.census.gov/library/publications/2024/demo/p60-284.html

Kernbach, J. M., & Staartjes, V. E. (2020). *Machine learning-based clinical prediction modeling: A practical guide for clinicians*. arXiv. https://doi.org/10.48550/arXiv.2006.15069

Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big data application in biomedical research and health care: A literature review. *Biomedical Informatics Insights, 16*, BII.S31559. https://doi.org/10.4137/BII.S31559

Mahajan, S., Caraballo, C., Lu, Y., Valero-Elizondo, J., Massey, D., Annapureddy, A. R., Roy, B., Riley, C., Murugiah, K., Onuma, O., Nunez-Smith, M., Forman, H. P., Nasir, K., & Krumholz, H. M. (2021). Trends in differences in health status and health care access and affordability by race and ethnicity in the United States, 1999–2018. *JAMA, 326*(7), 637–648. https://doi.org/10.1001/jama.2021.9168

Mukasa, K. (2026). AI-driven forecasting for Medicare risk score modernization. International Journal for Multidisciplinary Research, 8(1). https://doi.org/10.36948/ijfmr.2026.v08i01.66658

Ndugga, N., Hill, L., Rao, A., Pillai, A., & Artiga, S. (2025). *Key data on health and health care by race and ethnicity*. KFF. https://www.kff.org/racial-equity-and-health-policy/key-data-on-health-and-health-care-by-race-and-ethnicity/

Nkouaga, F. (2025). Socio-economic and insurance factors in preventive care: A two-stage weighted structural analysis. *Dialogues in Health, 7*, 100243. https://doi.org/10.1016/j.dialog.2025.100243

Nwoke, J. (2024). Healthcare data analytics and predictive modelling: Enhancing outcomes in resource allocation, disease prevalence and high-risk populations. *International Journal of Health Sciences, 7*(7), 1–35. https://doi.org/10.47941/ijhs.2245

Office of Disease Prevention and Health Promotion. (n.d.). *Social determinants of health*. U.S. Department of Health and Human Services. https://odphp.health.gov/healthypeople/priority-areas/social-determinants-health

Ogunsakin, O. L., & Mukasa, K. (2026). Navigating ethics and regulation in AI health cloud: Challenges and opportunities for United States healthcare. International Journal of Innovative Science and Research Technology, 11(1), 942–952. https://doi.org/10.38124/ijisrt/26jan367

Parsons, V. L., Moriarity, C., Jonas, K., Moore, T. F., Davis, K. E., & Tompkins, L. (2014). Design and estimation for the National Health Interview Survey, 2006–2015. *Vital and Health Statistics, Series 2, 165*, 1–53.

Penchansky, R., & Thomas, J. W. (1981). The concept of access: Definition and relationship to consumer satisfaction. *Medical Care, 19*(2), 127–140. https://doi.org/10.1097/00005650-198102000-00001

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research, 96*(1), 3–14. https://doi.org/10.1080/00220670209598786.

Raghupathi, W., & Raghupathi, V. (2018). An empirical study of chronic diseases in the United States: A visual analytics approach to public health. *International Journal of Environmental Research and Public Health, 15*(3), 431. https://doi.org/10.3390/ijerph15030431

Ratnapradipa, D., Brown, J., Jiang, L., & Qu, M. (2023). Where people live influences how they age: Evaluating social determinants of health and Medicare expenditures across geographic taxonomies. *Inquiry, 60*, 1–12. https://doi.org/10.1177/00469580231172505

Rojas, J.C., Rohweder, G., Guptill, J. *et al.* Predictive Analytics Programs at Large Healthcare Systems in the USA: a National Survey. *J GEN INTERN MED* **37**, 4015–4017 (2022). https://doi.org/10.1007/s11606-022-07517-1

Shi, L., Chen, C. C., Nie, X., Zhu, J., & Hu, R. (2014). Racial and socioeconomic disparities in access to primary care among people with chronic conditions. Journal of the American Board of Family Medicine : JABFM, 27(2), 189–198. https://doi.org/10.3122/jabfm.2014.02.130246

Snyder, J. E., Upton, R. D., Hassett, T. C., Lee, H., Nouri, Z., & Dill, M. (2023). Black Representation in the Primary Care Physician Workforce and Its Association With Population Life Expectancy and Mortality Rates in the US. JAMA network open, 6(4), e236687. https://doi.org/10.1001/jamanetworkopen.2023.6687

Tikkanen, R., & Abrams, M. (2020, January 30). U.S. Health Care from a Global Perspective, 2019: Higher Spending, Worse Outcomes? Commonwealth Fund. https://www.commonwealthfund.org/publications/issue-briefs/2020/jan/us-health-care-global-perspective-2019Waters, H., & Graf, M. (2018). *The costs of chronic disease in the United States*. Milken Institute.

Waters, H., & Graf, M. (2018). *The costs of chronic disease in the U.S.* Milken Institute. https://milkeninstitute.org/content-hub/research-and-reports/reports/costs-chronic-disease-us

Wesson, P., Hswen, Y., Valdes, G., Stojanovski, K., & Handley, M. A. (2022). Risks and opportunities to ensure equity in the application of big data research in public health. *Annual Review of Public Health, 43*, 1–18.

Williams, D. R., Lawrence, J. A., & Davis, B. A. (2019). Racism and health: Evidence and needed research. *Annual Review of Public Health, 40*, 105–125. https://doi.org/10.1146/annurev-publhealth-040218-043750

Wolfe, M. K., McDonald, N. C., & Holmes, G. M. (2020). Transportation barriers to health care in the United States: Findings from the National Health Interview Survey, 1997–2017. *American Journal of Public Health, 110*(6), 815–822. https://doi.org/10.2105/AJPH.2020.305579

World Health Organization. (2017). *Social determinants of health: Key facts*. https://www.who.int/news-room/fact-sheets/detail/social-determinants-of-health

Yearby, R. (2020). Structural racism and health disparities: Reconfiguring the social determinants of health framework to include the root cause, law. *Journal of Law, Medicine & Ethics, 48*(3), 530–545. https://doi.org/10.1177/1073110520958876

Zhao, Y., Wood, E. P., Mirin, N., Cook, S. H., & Chunara, R. (2021). Social determinants in machine learning cardiovascular disease prediction models: A systematic review. *American Journal of Preventive Medicine, 61*(4), 596–605. https://doi.org/10.1016/j.amepre.2021.04.016